DOCUMENT RESUME

ED 255 580 TM 850 227

AUTHOR Haladyna, Thomas M.; Downing, Steven M.
TITLE A Quantitative Review of Research on Multiple-Choice Item Writing.
SPONS AGENCY American Coll. Testing Program, Iowa City, Iowa. Research and Development Div.
PUB DATE Apr 85
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (69th, Chicago, IL, March 31-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Research; Elementary Secondary Education; Item Analysis; *Multiple Choice Tests; Research Reports; *Test Construction; *Test Items; Test Reliability; Test Validity; Textbooks
IDENTIFIERS Distractors (Tests); *Quantitative Research

ABSTRACT

In this paper 45 item-writing rules for multiple-choice tests presented in textbooks on educational measurement in a previous study are identified. The current study presents a quantitative review of the literature with respect to the empirical and theoretical evaluation of these principles of item-writing. Fifty-six studies that addressed at least one of the 45 item-writing rules were identified. Twenty-one of the rules have been studied empirically; 24 item-writing rules have no empirical basis. The optimal number of options was the most frequently studied rule, with 18 studies cited. The major generalization from these studies is that three options maximize test reliability and efficiency. Type-k items were evaluated in eight studies. Results suggest that compared to single-answer multiple-choice items, type-k items are more difficult, provide clues to some examinees, and decrease test reliability and efficiency. Eight other studies suggest some empirical basis for keeping the length of the keyed option about the same as other options. This review suggests that the majority of the common principles of multiple-choice item writing are not empirically based. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED255580

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A QUANTITATIVE REVIEW OF RESEARCH ON MULTIPLE-CHOICE ITEM WRITING

Thomas M. Haladyna and Steven M. Downing

The American College Testing Program
P.O. Box 168
Iowa City, Iowa 52243

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S. M. Downing
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 850 227

A paper presented at the annual meeting of the American Educational Research Association, Chicago, IL., April, 1985.

A Quantitative Review of Research on Multiple-Choice Item Writing

Thomas M. Haladyna and Steven M. Downing
The American College Testing Program
Iowa City, IA 52243

ABSTRACT

In a previous study the authors identified 45 item-writing rules for multiple-chioce tests, presented by authors of textbooks in educational measurement. The current study reports a quantitative review of the literature with respect to the empirical and theoretical evaluaton of these principles of item writing.

Fifty-six studies that addressed at least one of the 45 item-writing rules were identified. Twenty-one (47%) of the rules have been studied empirically; twenty-four (53%) item-writing rules have no empirical basis.

The optimal number of options was the most frequently studied rule with 18 studies cited. The major generalization from these studies is that three options maximize test reliability and efficiency.

Type-k items were evaluated in eight studies. Results suggest that compared to single-answer multiple-choice items, type-k items are more difficult, provide clues to some examinees, and decrease test reliability and efficiency. Eight other studies suggest some empirical basis for keeping the length of the keyed option about the same as other options. All other rules had six or fewer studies.

This review suggests that the majority of the common principles of multiple-choice item writing are not empirically based. Current item-writing practices remain more art than science.

A paper presented at the annual meeting of the American Educational Research Association, Chicago, IL., April 1985.

Most multiple-choice item writers receive initial instruction from any number of textbooks that deal with educational or psychological testing. The sum of knowledge about multiple-choice item writing is not found in any single reference but exists as lore passed down from generation to generation through these textbooks. Despite many advances in test theory in recent years, such as generalizability theory (Brennan, 1983), and item response theory (Lord, 1980), item writing has not yet advanced far as a science, although a number of theories of item writing have been proposed (Bormuth, 1970; Roid and Haladyna, 1982).

The present study is the second in a series of studies concerning multiple-choice item-writing practices. The objective in the first study (Haladyna & Downing, 1984) was to examine these textbooks and identify the core of knowledge about multiple-choice item writing. The objective in this second study is to examine the research base that supports item-writing practices as promulgated in these textbooks. The studies date from 1925 to 1984 and span a wide variety of test content, educational levels, test types, and, of course, item-writing practices. Quantitative methods were used in an effort to synthesize the results found in the studies. Before these are discussed, however, the Haladyna and Downing (1984) study will be reviewed as a means of presenting the basis for the present research.

An Analysis of Knowledge About Multiple-Choice Item Writing

Thirty-five textbooks that represent a wide range of perspectives and periods in educational testing have been identified (see Appendix A). Instructional statements were identified in these textbooks, and organized by six fundamental categories: (1) general item-writing advice--content concerns (2) general item-writing advice--construction, (3) item advice focusing on stem construction, (4) general advice focusing on option construction,

(5) advice focusing on construction of the correct option, and (6) advice focusing on the construction of the distracters (incorrect options).

The researchers identified which passages in each textbook discussed multiple-choice item writing, and classified all of the instructional statements contained in these passages. It was possible to construct an author-by-rule matrix and observe the number of instructional statements made by each textbook author, the frequency of occurrence of each rule across all textbooks, and the number of different rules that existed in the textbook literature.

Initially, 50 rules were identified. Upon closer examination, the list was refined to 45 rules, and these formed the basis for the present study. Table 1 summarizes the 45 rules according to the six categories previously discussed.

Of these rules, 14 were identified as appearing most frequently. Three of these rules could be considered general advice, eight as advice on option construction, and the remaining three suggestions on distracter construction. Interestingly, many of the most frequently used rules are the kind that are empirically testable (e.g., avoid the use of "none of the above"), rather than the type of rule that is based largely on common sense and is not easily empirically testable (e.g., "avoid items based on opinions" or "make a good transition from stem to option").

With respect to the frequency with which these rules are cited, Ebel (1979) led all other textbook authors. However, he cited only 58% of all rules. For other authors, this percentage of citation of all rules ranged downward to 20%.

Finally, the researchers identified the number of citations to research on multiple-choice item writing that each textbook contained. The number of citations ranged from zero to 24, with a median of 2.5.

Thus it would appear that a body of knowledge does exist for multiple-choice item writing. Most authors, however, do not appear to use the research literature to substantiate their advice. They may instead depend on what they have learned through courses they have taken, experiences in item writing, and other sources.

The present research study was a natural consequence of the first. The objective of this study, as mentioned earlier in this paper, was to explore the research basis for these instructional statements on item writing. More specifically, three questions were addressed in the present study:

1. How many studies deal with these item-writing rules?
2. Which of the item writing rules have been most often studied?
3. For the rules that have been most often studied, what conclusions can be drawn regarding their validity?

METHOD

Design of the Study

This review is quantitative in the sense that the number of studies reported in the literature and the frequency with which item-writing rules have been studied are its central foci. Further, results were evaluated in terms of ratings of effects rather than by other more subjective methods. This procedure is a middle ground between the more traditional review procedure and meta-analysis. The former type of research method is flawed by the problem of subjectivity. The latter requires a large number of studies and data that can be aggregated, neither of which could be obtained for this review. It was not possible, therefore, to use meta-analysis techniques.

Search Procedures

The search for research studies dealing with any of the 45 item-writing rules began with a computerized literature search on the topic "item writing." Each of the papers identified was reviewed and was either accepted or rejected for further consideration. Many papers dealt with theoretical approaches to item writing, such as those found in Roid and Haladyna (1982), and these were eliminated because these item-writing rules were not the concern of this review. References from those papers included in this study were examined for leads to other studies. This process assured that most relevant research was identified and included in the present study.

Method for Classifying Studies

A coding sheet was used to classify each study. The types of information coded included (1) sample size, (2) test length, (3) type of test (i.e., standardized achievement, classroom achievement, or aptitude-ability), (4) approximate educational level of the examinees, (5) rules studied, (6) methodological problems, and (7) a rating for each criterion involving each rule.

Results were evaluated on the basis of six criteria typically used in these studies: (1) item difficulty, (2) item discrimination, (3) reliability, (4) validity, (5) efficiency (the time it takes to complete a test), and (6) test score variance.

Both authors of the present study validated the rating form by individually rating five studies and comparing ratings. The findings showed concurrence, so the balance of these papers were divided for review. In the course of synthesizing these studies, all studies were reviewed again, and discrepancies in classification were resolved through mutual agreement.

Analysis of Results

To answer the study question about how many studies address item-writing rules, the number of papers rated was counted. This simply provided an overall measure of what kind of attention item-writing research has received in the empirical literature.

The second question dealt with the frequency with which each item-writing rule had been studied. A frequency distribution was created for the 45 rules.

It was more difficult to draw conclusions about the validity of rules, which was the point of the third research question. All studies with a frequency of two or more were subjected to additional review to determine if any consensus could be found among the studies. The intent was to discover if the rule had analytical or theoretical support as well as empirical support. For some rules, it was possible to synthesize all studies that discussed the rule.

RESULTS AND DISCUSSION

To answer the first question of this study, 56 studies were identified that addressed at least one of the 45 item-writing rules. These studies varied widely with respect to types of tests, test lengths, sample sizes, and educational levels of samples. All of these studies were published between 1925 and 1984. As the availability of computers improved studies in the 1970s and 1980s, the method of statistical analysis changed significantly. Nonetheless, some of the best designed and most comprehensive studies were completed in the 1920s.

Table 2 provides the frequency distribution of the rules studied most often. As shown there, only the rule dealing with the optimal number of options has received major attention, while five other rules have received

moderate attention. All other rules were cited four or fewer times; seven rules received only one citation, while 24 rules were not cited at all. The balance of this section will be devoted to discussions of the research on the most frequently studied rules.

Rule 26: Use three, four, or five options for an item.

Studies of the ideal number of options can be divided into two discrete groups: (a) theoretical and analytical, or (b) empirical. Each will be discussed in turn.

The earliest study of option number was by Lord (1944), who developed a formula for predicting changes in reliability as a function of the number of options added to a multiple-choice item. Lord's data suggest that three-option items are optimal. Tversky (1964) developed three criteria (discriminability, power, and information of a test) to evaluate the number of choices in a multiple-choice item. He concluded that (1964, p 390):

> Whenever the amount of time spent on the test is proportional to its total number of alternatives, the use of three alternatives at each choice point will maximize the amount of information obtained per time unit.

This finding has been supported in subsequent studies by Ebel (1969), Grier (1975; 1976), and by Lord (1977). Lord's study (1977) is most informative about the point at which three-option test items are most effective. Using item response theory, Lord presents item efficiency curves to show that the three-option item provides maximum information in the mid-range of the score scale, while the true-false item provides most information for high-scoring examinees and the four-option or five-option item provides the most information for low-scoring examinees. This is an interesting and important observation that takes into account the prominence of guessing among low-scoring examinees. (And, of course, the four-option or five-option item offers more protection against guessing.) Because high-scoring examinees are

less likely to guess, two options are sufficient for them. For most examinees, providing three options appears to be optimal.

Test reliability is the only index of overall test quality compared in all empirically based studies on the optimal number of options for multiple-choice items. Other characteristics of items and tests compared in these studies were item difficulty and discrimination, validity, test score variance, and efficiency.

Reliability. Table 3 presents reliability coefficients for the ten studies that report reliability coefficients for tests with various numbers of options. These reliabilities are computed by different formulas under the conditions of various test lengths, sample sizes, educational levels, and test content.

Table 3 shows that reliability is, in general, a monotonically increasing function of the number of options, but that the incremental gain in reliability is small when more than three options are used. The authors of these studies conclude that when efficiency is taken into account--the extra effort needed to create additional options, and the extra time needed for students to respond to longer items--either three or four options maximize reliability.

Efficiency. Several studies examine efficiency for various numbers of options. Efficiency is variously defined from "absolute time to complete" to the relative efficiency of information bits gained per unit of time. For example, Williams and Ebel (1957) conclude that two or three options are most efficient, but in an earlier study, Ruch and Stoddard (1925) state that two or five options maximize efficiency. In general, these studies conclude that three or four options are most efficient.

Item difficulty and discrimination. Several studies report item difficulty and discrimination, and the results are mixed and contradictory. For example, Charles (1926) reports that five option items are the most difficult and two option items are the least difficult. Costin (1972) reports no differences in item difficulty or discrimination between three and four options. Parker and Somers (1983) show no differences in difficulty between four and five option items. Straton and Catts (1980) compare two, three, and four options and report that three-option and four-option items are nearly equal in difficulty, but that three-option items discriminate better than four-option items.

In summary, the relationship of the number of options to test reliability is the most frequently studied item writing practice. In general, test reliability is shown to vary directly with the number of options from two to five. However, the incremental gain in test reliability when a fourth or fifth option is added is very small. In general, these studies show that test efficiency is to be maximized by three or four options. Item difficulty and discrimination show mixed results for two to five options and no conclusions are warranted. Validity was studied in only one study (Ruch & Stoddard, 1925), which showed that five options increased criterion-related validity.

Rule 19: Avoid type-k items.

Of the studies involving type-k items, all but one involved comparisons with the type-x (multiple true-false) format. Therefore, comparisons with conventional multiple-choice were limited to two studies, but the other studies, involving the x-type items, provide additional insights about the type-k.

Parker and Somers (1983) compared the type-k format with four-option and five-option multiple-choice items, and found the type-k format more difficult

as well as less reliable than the other two formats. Hughes and Trimble (1965) compared a precursor of the type-k format where the option "both are correct" is used. Their findings indicated higher reliability for the "both" option as well as higher variance of test scores. Difficulty was unaffected. In a replication, they found that the "both" option, when compared to a conventional format, increased reliability and variance and also produced more difficult items. A second replication yielded results similar to those of the first replication--more difficulty and greater reliability. No effect on item discrimination was detected. This contradicts the finding of higher reliability for this format, because item discrimination and reliability are functionally related.

The results of the studies involving type-k and type-x items are somewhat mixed. Further, this research is somewhat confounded because scoring systems for type-x items vary significantly, and because the chance levels for type-k and type-x items are not the same, which makes test scale comparisons somewhat problematic.

Regarding item difficulty, the results of the studies are mixed, perhaps owing to the variety of scoring methods for type-x formats. Albanese, Kent, and Whitney (1979), Harasym, Norris and Lorscheider (1980) and Kolstad, Briggs, Bryant, and Kolstad (1983) report the type-k format produced easier items, while Albanese, Kent, and Whitney (1977) found the opposite. None of these studies addressed item discrimination. Three studies (Albanese et al., 1977; Albanese et al., 1979; and Harasym et al., 1980) all reported lower reliability with the type-k format, while only Hill and Woods (1976) report no differences between type-k and type-x formats. With respect to validity, only two studies (Hill & Woods, 1976; Albanese et al., 1979) reported no differences.

In summarizing these results, it must be noted that the paucity of studies comparing type-k with conventional multiple-choice items is a serious limitation. However, these studies present some strong arguments against type-k formats.

1. In most circumstances, this format seems to produce more difficult items. Although increased difficulty need not be a problem, it can be if not taken into account when a test with both type-k items and items that have other formats is assembled.

2. The suspicion that type-k items provide clues is shared by Harasym et al., (1980) and Albanese (1982) who offer evidence in support of this belief. It appears that, unlike knowledge about the truth of a primary option, knowledge about the falsity of an option helps eliminate the secondary choices in the type-k format. It therefore seems very possible that type-k items help clue examinees, particularly low-scoring ones. However, this needs to be more extensively studied.

3. It is clear that type-k items are less reliable in most instances.

4. Perhaps the most compelling reason for rejecting the type-k item is that it is more inefficient to construct and more laborious to read. More of the conventional multiple-choice items than the type-k items can be given per unit of time.

5. Finally, the finding of Hill and Woods (1976) that students prefer x-type over k-type cannot be ignored. Although hardly a sufficient condition, face validity is certainly necessary in the choice of a test format.

Rule 29: Keep the length of options fairly consistent.

Eight studies concerning the effect of presenting the keyed option as the longest alternative were reviewed. All of these studies evaluated the effect of the key being the longest option on item difficulty, while some studies

also evaluated the effect of this flaw on item discrimination, test reliability, and concurrent validity.

Board and Whitney (1972) found that length of keyed options made no overall difference in test difficulty. However, they also found that less able students tended to use the clue of longer keyed options more than abler students. Both test reliability and concurrent validity were decreased by the length flaw. In the design of this study, course final examination score was used as a blocking variable.

Chase (1964) also found that the length of the correct option had no effect on difficulty, but concluded that the response set to select the longest option interacts with item difficulty. For more difficult items, then, students tend to use the length clue, but for easier items they do not.

All other studies reviewed concluded that the length flaw produced easier items. Jones and Kaufman (1975), in a study of response set, found that higher-scoring students use the length clue more than do lower-scoring students. An internal total test score criterion was used to block high and low-scoring students in this study.

Evans (1984) and Strang (1977) found longer keyed options to be easier than shorter keyed options. Dunn and Goldstein (1959) and McMorris et al. (1972) concluded that the length clue made items easier, but had no effect on reliability and validity. Weiten (1984) also found the longer keyed alternative to be easier, but there was no effect on item discrimination, test reliability, or validity.

In summary, most studies conclude that the use of long correct options makes items easier. In the only two studies that note no such effect, student ability was used as a blocking variable with contradictory results. The difference in measures of student ability used in these two studies may

account for the contradictory findings. This item writing flaw lowers test reliability and concurrent validity in only one of the eight studies reviewed.

Rule 30: Avoid the use of "none of the above".

A total of six studies that discussed the use of "none of the above" were reviewed. These studies examined the effect of this option on item difficulty and discrimination and on test reliability and validity.

Schmeiser and Whitney (1975b), in an extensive study of the use of the "none of the above" option, found that the effect on difficulty and discrimination on tests of different subject matter was mixed. According to their findings, test reliability and validity were slightly decreased by the use of this option.

Wesman and Bennett (1946) observed no effect on item difficulty and a mixed result on item discrimination. The data from this study are, however, difficult to interpret. A mixed effect on test difficulty and item discrimination were also reported by Williamson and Hopkins (1967). However, examination reliability was lower for this type of option.

Studies by Dudycha and Carpenter (1973), Hughes and Trimble (1965), and Rinland (1960b) concluded that "none of the above" increased item difficulty. Dudycha and Carpenter (1973) and Hughes and Trimble (1965) also found lower test reliability, but Rinland (1960b) did not evaluate the effect of this practice on reliability.

In summary, no conclusion can be reached about the effect of "none of the above" on item difficulty. Three studies found that this option increased item difficulty, but three studies reported no effect or mixed results. Three studies found that the use of this option lowers test reliability. There is no consensus about the effect on item discrimination. Only one study

evaluated concurrent validity effects (Schmeiser & Whitney, 1975b), and it found validity slightly decreased by use of "none of the above."

Rule 17: State the stem in either a question form or a sentence form.

This rule has received moderate attention in the research literature. Six studies are cited (Board & Whitney, 1972; Dudycha & Carpenter, 1973; Dunn & Goldstein, 1959; Schrock & Mueller, 1982; Schmeiser & Whitney, 1975a; 1975b). As in most other instances, the test lengths, test types, educational levels of examinees, test content, and other factors vary significantly across these papers. Despite this variability, some definite trends in findings about this rule can be reported. In four of the six studies, incomplete stems were found to be more difficult. While the practical magnitude of this difference is small, it could affect test assembly, because a preponderance of incomplete stems will produce a systematically more difficult test.

Typically, discrimination was unaffected by completeness of the stem. Reliability and validity appear to be slightly affected, but this result may be due to the way in which discrimination was calculated. When the upper 27% to lower 27% index is used instead of the point-biserial, discrimination may not be accurately estimated, since the former is only an approximation of the more desirable latter. Since discrimination and reliability are functionally related, significant differences in discrimination logically lead to significant differences in reliability.

Thus, based on this limited set of studies, it is possible to draw the preliminary conclusion that the incomplete stem is a less effective item-writing strategy than the question format. While the differences between the two stem types are slight, the replication of findings builds support for this conclusion in the absence of further studies.

Rule 35: Balance the key; that is, make sure the correct option is found an equal number of times in each option position, if possible.

Six studies concerning key balancing were reviewed in the present research. The results of these studies were mixed. Four studies reported that the position of the keyed response affected item difficulty and two studies reported the opposite.

Ace and Dawis (1973), Jones and Kaufman (1975), Evans (1984), and McNamara and Weitzman (1945), report that the position of the key is related to item difficulty. Ace and Dawis (1973) found that the fifth position for the keyed response was the most difficult for examinees and the third position was next most difficult.

Both Marc[illegible]963) and Wilber (1966) report that there is no evidence of a positional response set or a relationship between position of the key and difficulty.

Rule 20: Don't clue through grammatical errors.

This rule refers to the inadvertent use of incorrect grammar to clue examinees to the correct option. Only four studies can be reported which have studied the validity of this rule.

Evans (1984) reported that grammatical cluing made items easier and increased the variability of the test scores. McMorris, Brown, Snyder, and Pruzek (1972) found that this fault made items easier, but no effect was noted on reliability or validity. Weiten (1984) found that difficulty and discrimination were not affected by grammatical inconsistency. Interestingly, the results for reliability were mixed, and the results for validity were inconclusive. Huntley and Plake (1980) found no support for cluing through grammatical inconsistency.

Nevertheless, it seems sensible to avoid grammatical error, just to support the face validity of the test. In the absence of more conclusive

empirical evidence, the rule should stand on the grounds that grammatical clues detract from face validity.

Rule 16: Avoid window dressing in the stem.

The effect of window dressing--extraneous material--in the stem of the item was investigated in four of the studies reviewed.

Rimland (1960a) found that window dressing decreased test reliability, discrimination, variance, and concurrent validity. Schrock and Mueller (1982) reported that window dressing made test items more difficult and took students longer to complete than items without window dressing. However, this item flaw did not affect test reliability.

Board and Whitney (1972) found that less able students performed better on items with window dressing than more able students. There was no overall effect on mean test difficulty, but a decrease in test reliability was reported. However, Schmeiser and Whitney (1975a) reported little or no effect of window dressing on item difficulty or test reliability.

In summary, three of the four studies reviewed suggest that window dressing has an adverse effect on at least some students.

Rule 24: Don't leave blanks in the middle of the stem.

This rule is similar to the rule about using a question stem rather than a stem that is an incomplete statement. The rule arose from verbal analogy items. so it has limited applicability, but the effect that leaving blanks in the middle of the stem sentence has on item and test characteristics may be of interest. Silverstein and McClain (1963) were among the first to examine the effect of blanks in items, although they allude to a study by Campbell (1961) in which a design flaw makes the results questionable. Silverstein and McClain (1963) found no effects when the blank was systematically varied in the stem. Ace and Dawis (1973) describe the dispute between Campbell and his

adversaries and offer partial support for both sides. Changes in the structure of the analogy did not change difficulty, but the interaction of this change and the position of the correct response did appear to affect difficulty.

Schrock and Mueller (1982) offer the only study that addresses this rule as it applies to items not based on analogies. Their findings seem to suggest virtually no effect on difficulty, test score variance, or response time but mixed results were reported for reliability.

The findings from these reports suggest that this rule is still strongly in need of further study. However, there does not seem to be any harm in leaving a blank in the middle of the stem. Until more evidence is marshalled to support it, the rule appears to have questionable validity.

Rule 38: Use plausible distracters.

This rule, like several others, appears to be based on common sense; empirical testing seems hardly necessary. Yet three very different studies discuss its applicability.

The first of these, by Weiten (1984), compared plausible and implausible options, and found that flawed items were less difficult, but not less discriminating. No differences were observed for reliability or validity, since the variance of test scores was maintained so that the testwiseness clues in these implausible distracters assisted all ability levels of examinees equally.

Smith (1982) used a very small sample of students and items to examine the tendency for students to determine the right answer by using a teachable strategy. Smith concludes that test-taking may be a learned skill, and that learning the skill may affect test scores. If distracters are written as variations of correct answers, as Smith contends, then convergence theory may

explain the development of testwiseness and may indicate that test scores are artificially inflated if distracters are plausible.

The third study, by Owens, Hanna, and Coppedge (1970), compares three methods of generating or selecting distracters: the judgmental method, the frequency method, and the discrimination method. The judgmental method, in which the item writer inverts the most plausible distracters, was directly compared to the frequency method, in which the actual responses that students made to open-ended questions were tallied and those written most frequently were used in subsequent multiple-choice tests. Results were mixed, at least with respect to reliability and validity. Difficulty and test score variation did not seem to be affected.

On the surface, it seems obvious that implausible distracters are not desirable. Yet two of the three studies provide compelling evidence that plausible distracters may be more easily eliminated by testwise examinees. The study by Owens et al. (1970) suggests that distracters for a test should be field tested and that the distracters should be chosen because they have negative discrimination and negative item characteristic curves.

Rule 40: Don't use distracters that clue testwise examinees.

Sarnecki (1979) has presented a very complete analysis of testwiseness. Testwiseness is an examinee characteristic and thus outside the scope of this review; however, some elements of item writing are influenced by testwiseness. Only three studies that discuss cluing answers by violating item-writing principles other than the rule of grammatical consistency were identified.

Each of the three studies focus on the repeating of a word or phrase from the stem in the correct option, which is a testwiseness clue. McMorris et al. (1972) and Weiten (1984) found that only item difficulty is affected by such

cluing. While Pyrczak (1973) did not replicate these findings, he did find that testwiseness could be taught and that some students could increase scores after training.

Despite the similarity in the design of these three studies, the rule discussed appears logically sound. The use of specific determiners (e.g., "always" and "never"), the use of cognates in the stem and correct option, and the use of ridiculous options should provide unfair advantage to testwise students. Thus the rule should be supported on grounds of prudence, but should be interpreted in light of the findings of Smith (1982) and Weiten (1984), discussed for Rule 38.

Rule 37: Use common errors of students for distracters.

This rule was briefly mentioned in conjunction with the study by Owens et al. (1970) supporting Rule 36. Their method for generating distracters was to use a completion format and have students respond. The errors that appeared most frequently were the bases for constructing distracters and produced good results according to their study. To take this principle a step further, student errors might also be evaluated in terms of their discriminating power; such distracters should have negative discrimination and negative item characteristic curves. The study by Powell and Isbister (1974) is unique in this review and worthy of more extensive attention. It examined the response patterns inherent in correct and incorrect answers, challenged the assumption that no useful information is available in wrong answers. This work suggests an interesting proposition that has received recent attention in other, more theoretical discussion of item writing: that items should have diagnostic distracters that provide information that not only increases test reliability (Haladyna, 1984) but also permits diagnostic instruction (Roid, 1984).

Rule 31: Avoid the use of "all of the above."

While most textbook authors recommend against using the "all of the above" option, only two studies can be reported here that address this rule. Dudycha and Carpenter (1973) report that use of "all of the above" makes items more difficult and less discriminating. Hughes and Trimble (1965) report that items that use the option "both of the above," described earlier as a precursor to the type-k format, are more difficult, but that both variance and reliability appear to be increased. These findings contradict those of Dydycha and Carpenter (1973).

In light of this disagreement, it is difficult to evaluate this rule. Authors and teachers are cautioned against recommending such item-writing practices without more experience or data to support such a rule.

Rule 32: Use the option "I don't know."

This option is intended to reduce the incidence of examinee guessing. Sanderson (1973) examined "don't know" in a clinical education setting and found that there was a slight distortion of scores by those using this option. Sherman (1974) examined National Assessment of Educational Progress data and found differences according to age, region, ethnic background, and even personality, in response patterns to this option. These findings are particularly impressive since these data are a national probability sample representing the entire United States.

Although only two studies have examined this rule, the evidence appears overwhelmingly in favor of rejecting its validity. Although it is meant to reduce guessing, guessing is confounded and testwiseness is rewarded. It is thus difficult to justify the use of an "I don't know" option.

Other Findings

Seven other rules received only one citation in an empirical study. The findings are presented in Table 4. In this table, the author and rule are identified and a box score is used to determine whether the rule is supported on various grounds such as difficulty, discrimination, reliability, and validity. These findings are presented here for completeness but are not discussed further because only one study has been identified for each rule. The remaining 24 rules received no attention.

CONCLUSIONS

Only 56 studies were found to bear directly on the validity of 45 item-writing rules: testimony enough that there has not been sufficient research to support most of them, although common sense and face validity suggest that many of the rules are legitimate.

The frequency with which rules have been empirically tested is directly related to the number of studies. Many of these studies address more than one rule, but few rules have been studied more than four times, and many rules are substantiated by little or no empirical research.

The optimal number of options that a multiple-choice item should have has received considerable attention. Empirical research supports theoretical study in indicating that three options achieve the optimal balance between reliability and efficiency. It is suprising, considering the evidence, that virtually all authors of textbooks favor multiple-choice items with four or five options and that nearly all standardized tests use more than three options.

The other rules do not have a firm foundation in research. Further study of the validity of item-writing rules is necessary. The paragraphs that follow suggest some fruitful areas for exploration.

1. It is desirable to find methods to improve the development of items that measure higher-level thinking. Few of the proposals in textbooks and other sources (e.g., Miller, Williams, and Haladyna, 1978), have met with practical success.

2. The research to date, particularly that of Lord (1977), indicates that item performance improves when distracters have negative item characteristic curves (i.e., negative discrimination). Ideally, distracters should have a diagnostic value. When a student selects a distracter, some valuable corrective teaching should be possible. A procedure like the answer-until-correct is a step in this direction and may prove to be a rewarding area for research.

3. The large number of rules yet unstudied provides a source for future research. Item writers need to know the merits or demerits of the "all" option, the "none" option, and the "don't know" option.

Methodological Concerns

Many studies reviewed for this paper are flawed. Further studies on item-writing rules must, to be of value, have a sound experimental design. Each of the factors under consideration must be well defined and completely tested via main effects and interactions. The samples of items and examinees must be sufficient to maintain a reasonable power for statistical tests.

Item difficulty and test difficulty have been vastly overemphasized in studies of item-writing rules. The effects of an item-writing practice on discrimination, reliability, validity, and efficiency are much more important and merit more attention. IRT methods may provide important insights to the effects of item-writing practices on test characteristics.

It is imperative that studies report the basic data used for analysis. Means and standard deviations are vital if the results of the study are to be properly interpreted.

Statistical tests are only the beginning of an analysis. The researcher should routinely report the effect size so that a standard can be used to evaluate a result. For large samples, virtually the smallest, most trivial difference is statistically significant. For small samples, a very large difference may be statistically insignificant.

Finally, it seems appropriate that item-writing practices should be based on item-writing theories, such as those suggested by Bormuth (1970) and Hively (1974). In addition to aiding in the definition of the construct to be measured (a necessary condition for the desirable construct validity), these theories also provide the bases for empirical research on the development of multiple-choice items.

Table 1

Coding System for
Classifying Instructional Statements on
Multiple-Choice Item Writing

General Item-Writing Advice

1. Avoid textbook, verbatim phrasing of items.
2. Avoid trick questions.
3. Avoid opinion-based items.
4. Base each item on an educational objective.
5. Use types of items that elicit higher-level thinking (various authors give examples and specific advice).
6. Test for important facts and knowledge.
7. Avoid items which require overspecific knowledge.

General Advice

8. Minimize examinee reading by limiting item length.
9. Use good grammar consistently, making sure that the item and the options agree grammatically.
10. Focus on a single, clearly defined problem in phrasing the question.
11. Consider vocabulary when phrasing the item; keep it appropriate for the intended audience.
12. Allow sufficient time for the development, review, and revision of the item.
13. Avoid interdependence of items or avoid allowing one item to cue another.
14. Format the item either horizontally or vertically.

Item Advice Focusing on Stem Construction

15. Ensure that the directions in the item stem are clear and that wording lets the examinee know what is being tested.
16. Avoid window dressing (extraneous materials) in the stem.
17. State the stem in either a question form or a sentence form with options completing the stem.
18. Use either the best answer or correct answer format.
19. Avoid type-k items, i.e., items that list a series of statements and then provide combinations of these statements as options.
20. Don't clue the correct response through a grammatical error.
21. Word the stem positively; avoid negatives.
22. Make a good transition from the stem to the options.
23. Include the central idea and most of the text of the item in the stem.
24. Stems should be left open at the end; don't leave blanks in the middle of the stem that refer to options.

Item Advice Focusing on Option Construction

General Advice

25. Items with different numbers of options may appear on the same test.
26. Use three, four or five options for an item.
27. Keep a logical order to options; if quantitative, keep options in ascending or descending order.
28. Keep options independent from one another.
29. Keep the length of options fairly consistent.
30. Avoid the use of "none of the above".
31. Avoid the use of "all of the above."
32. Use the option "I don't know."
33. Keep options homogenous in content and grammatical structure.
34. Phrase options positively, not negatively.

Correct Option

35. Balance the key; that is, make sure the correct option is found an equal number of times in each option position, if possible.
36. Make sure there is one and only one correct option.

Distracters

37. Incorporate common errors of students in developing distracters; anticipate what distracter is most likely to attract unprepared examinees.
38. Avoid illogical distracters; use plausible distracters.
39. Avoid specific determiners (e.g., never, always) in distracters.
40. Avoid distracters that can clue testwise examinees.
41. Avoid technically phrased distracters.
42. Use incorrect paraphrases as distracters.
43. Use familiar-looking but incorrect statements as distracters.
44. Use true statements that do not correctly answer the question as distracters.
45. Use irrelevant clues for distracters.

Table 2

Frequency of Studies for Each Item Writing Rule

Number	Rule	Frequency
26	Use three, four, or five options for an item.	18
19	Avoid type-k items.	8
29	Keep the length of options fairly consistent.	8
30	Avoid the use of "none of the above."	6
17	State the stem in either a question form or a sentence form.	6
35	Balance the key.	6
20	Don't clue through grammatical errors.	4
16	Avoid window dressing in the stem.	4
24	Don't leave blanks in the middle of the stem.	4
38	Use plausible distracters.	3
40	Don't use distracters that clue testwise examinees.	3
37	Use common errors of students for distracters.	2
31	Avoid the use of "all of the above."	2
32	Use the option "I don't know."	2
	7 other rules had one study each.	1
	24 other rules had no studies cited.	0

Table 3

Reliability Coefficients for Items
of Two to Five Alternatives

	Number of Options			
	2	3	4	5
Charles (1926)	.477	.624	-	.680
Costin (1970)	-	.560	.500	-
Costin (1972)	-	.750	.780	-
Parker & Somers (1983)	-	-	.532	.562
Ramos & Stern (1973)	-	-	.860	.890
Ruch & Charles (1928)	.477	.624	-	.680
Ruch & Stoddard (1925)	.737	.598	-	.796
Straton & Catts (1980)	.470	.730	.680	-
Wakefield (1958)	.86[illegible]	.890	.920	.930
Williams & Ebel (1957)	.954	.945	.945	-

Table 4

Effects of Rules Evaluated in Single Studies

		Test Characteristics[a]			
	Rule	Difficulty	Discrimination	Reliability	Validity
Baker (1971)	4	+/-		0	
Dudycha & Carpenter (1973)	21	+	+/1		
Dunn & Goldstein (1959)	9	+			-
Kolstad, Goaz, & Kolstad (1982)	36	0		0	
Strang (1977)	41	+			
Strang (1977)	43	+			
Terranova (1969)	34	+		0	

Note: Interpretation of symbols

a. Positive effect +
Negative effect -
Inconclusive effect 0
Mixed effect +/-

References

Ace, M. C., & Dawis, R. V. (1973). Item structure as a determinant of item difficulty in verbal analogies. Educational and Psychological Measurement, 33, 143-149.

Albanese, M. A., (1982). Multiple-choice items with combinations of correct responses. Evaluation and the Health Professions, 5, 218-226.

Albanese, M. A., Kent, T. H., & Whitney, D. R. (1977). A comparison of the difficulty, reliability, and validity of complex multiple choice, multiple response and multiple true-false items (Research Report No. 95) Iowa City, Iowa: University of Iowa, Learning Resources Unit, College of Medicine and Evaluation and Examination Services.

Albanese, M. A., Kent, T. H. & Whitney, D. R. (1979). Cluing in multiple-choice test items with combinations of correct responses. Journal of Medical Education, 54, 948-950.

Baker, E. L. (1971). The effects of manipulated item writing constraints on the homogeneity of test items. Journal of Educational Measurement, 8, 305-309.

Board, C. B., & Whitney, D. R. (1972). The effect of selected poor item-writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 9, 225-233.

Bormuth, J. R. (1970). On the theory of achievement test items. Chicago, IL: University of Chicago Press.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Campbell, A. C. (1961). Some determinates of non-verbal classification items. Educational and Psychological Measurement, 21, 899-913.

Charles, J. W. (1926). A comparison of five types of objective tests in elementary psychology. Unpublished doctoral dissertation, State University of Iowa.

Chase, C. (1964). Relative length of options and response set in multiple choice items. Journal of Educational Measurement, 1, 58.

Costin, F. (1970). The optimal number of alternatives in multiple choice achievement tests: Some empirical evidence for a mathematical proof. Educational and Psychological Measurement, 30, 353-358.

Costin, F. (1972). Three-choice versus four-choice items: Implications for reliability and validity of objective achievement tests. Educational and Psychological Measurement, 32, 353-358.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item formats on item discrimination and difficulty. Journal of Applied Psychology, 58, 116-121.

Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and

reliability as a function of selected multiple-choice item construction principles. Educational and Psychological Measurement, 19, 171-179.

Ebel, R. L. (1979). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Ebel, R. L. (1982). Proposed solutions to two problems of test construction. Journal of Educational Measurement, 19, 267-278.

Ebel, R. L. (1969). Expected reliability as a function of choices per item. Educational and Psychological Measurement, 29, 565-570.

Evans, W. (1984). Testwiseness: An examination of cue-using strategies. Journal of Experimental Education, 52, 141-144.

Grier, B. (1976). The optimal number of alternatives at a choice point with travel time considered. Journal of Mathematical Psychology, 14, 91-97.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. Journal of Educational Measurement, 12, 109-113.

Haladyna, T. M. (1984). Increasing information from multiple-choice test items. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Haladyna, T. M., & Downing, S. M. (1984). An analysis of knowledge about multiple-choice test item writing. Paper presented at the annual meeting of the American Psychological Association, Toronto, Canada.

Harasym, P. H., Norris, D. A., & Lorscheider, F. L. (1980). Evaluating student multiple-choice responses: Effect of coded and free formats. Evaluation and the Health Professions, 3, 63-84.

Hill, G. C., & Woods, G. T. (1976). Multiple true-false questions. Education in Chemistry, 11, 86-87.

Hivels, W. (Ed.) (1974). Domain-referenced testing. Englewood Cliffs, NJ: Educational Technology Publications.

Hughes, H. H., & Trimble, W. E. (1965). The use of complex alternatives in multiple-choice items. Educational and Psychological Measurement, 25, 117-126.

Huntley, R. M., & Plake, B. S. (1980). Effect of selected item-writing practices on test performance: Can relevant grammatical clues result in flawed items? Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Jones, P. D., & Kaufman, G. G. (1975). The differential formation of response sets by specific determiners. Educational and Psychological Measurement, 35, 821-833.

Kolstad, R. K., Briggs, L. D., Bryant, B. B., & Kolstad, R. A. (1983). Complex multiple-choice items fail to measure achievement. Journal of Research and Development in Education, 17, 7-11.

Kolstad, R., Goaz, R., & Kolstad, R. (1982). Nonrestricted multiple-choice examination items. Journal of Dental Education, 46, 485-488.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1977). Optimal number of choices per item: A comparison of four approaches. Journal of Educational Measurement, 14, 33-38.

Lord, F. M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. Journal of Educational Measurement, 35, 175-180.

Marcus, A. (1963). The effect of correct response location on the difficulty level of multiple-choice questions. Journal of Applied Psychology, 47, 48-51.

Miller, H. G., Williams, R. G., & Haladyna, T. M. (1978). Beyond facts: objective ways to measure thinking. Englewood Cliffs: NJ: Educational Technology Publications.

McMorris, R. F., Brown, J. A., Snyder, G. W., & Pruzek, R. M. (1972). Effects of violating item construction principles. Journal of Educational Measurement, 9, 48-51.

McNamara, W. J., & Weitzman, E. (1945). The effect of choice placement on the difficulty of multiple-choice questions. Journal of Educational Psychology, 36 103-113.

Owens, R. E., Hanna, G. S., & Coppedge, F. L. (1970). Comparison of multiple-choice tests using different types of distracter selection techniques. Journal of Educational Measurement, 7, 87-90.

Parker, C. S., & Somers, J. E. (1983). A comparison of the difficulty and reliability of type K and one-best-response test items. A paper presented at the annual meeting of the Iowa Evaluation and Research Association, Des Moines, IA.

Powell, J. C., & Isbister, A. G. (1974). A comparison between right and wrong answers on a multiple choice test. Educational and Psychological Measurement, 34, 499-509.

Pyrczak, F. (1973). Use of similarities between stems and keyed choices in multiple-choice items. Paper presented at the annual meeting of the National Council for Measurement in Education.

Ramos, R. A., & Stern, J. (1973). Item behavior associated with changes in the number of alternatives in multiple-choice items. Journal of Educational Measurement, 10, 305-310.

Rimland, B. (1960a). The effect of including extraneous numerical information in a test of arithmetic reasoning. Educational and Psychological Measurement, 20, 787-794.

Rimland, B. (1960b). The effects of varying time limits and of using "right answer not given" in experimental forms of the U. S. Navy Arithmetic Test. Educational and Psychological Measurement, 20, 533-539.

Roid, G. H. (1984). Generating test items. In R. A. Berk (Ed.), Guide to criterion-referenced test construction. (pp. 49-77). Baltimore, MD: Johns Hopkins.

Roid, G. H., & Haladyna, T. M. (1982). A technology for test item writing. New York, NY: Academic Press.

Ruch, G. M., & Charles, J. W. (1928). A comparison of five types of objective tests in elementary psychology. Journal of Applied Psychology, 12, 398-404.

Ruch, G. M., & Stoddard, G. D. (1925). Comparative reliabilities of objective examinations. Journal of Educational Psychology, 16, 89-103.

Sanderson, P. H. (1973). The 'don't know' option in MCQ examinations. British Journal of Medical Education, 7, 25-29.

Sarnecki, R. E. (1979). An examination of testwiseness in the cognitive domain. Review of Educational Research, 49, 252-279.

Schmeiser, C. B., & Whitney, D. R. (1975a). Effect of two selected item-writing practices on test difficulty, discrimination, and reliability. Journal of Experimental Education, 1975, 43, 30-34.

Schmeiser, C. B., & Whitney, D. R. (1975b). The effect of incomplete stems and "none of the above" foils on test and item characteristics. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

Schrock, T. J., and Mueller, D. J. (1982). Effects of violating three multiple-choice item construction principles. The Journal of Educational Research, 75, 314-318.

Sherman, S. W. (1974). Multiple choice test bias uncovered by use of an "I don't know" alternative. Paper presented at the annual meeting of the National Academy of Sciences.

Silverstein, A. B., & McLain, R. E. (1963). Note on the internal structure of verbal analogy items. Psychological Reports, 12, 434.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. Journal of Educational Measurement, 19, 211-220.

Strang, H. R. (1977). The effects of technical and unfamiliar options on guessing on multiple-choice test items. Journal of Educational Measurement, 14, 253-260.

Straton, R. G., & Catts, R. M. (1980). A comparison of two, three, and four-choice item tests given a fixed total number of choices. Educational and Psychological Measurement, 40, 357-365.

Terranova, C. (1969). The effects of negative stems in multiple-choice test items. Dissertation Abstracts International, 30, 2390A.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. Journal of Mathematical Psychology, 2, 386-391.

Wakefield, J. A. (1958). Does the fifth choice strengthen a test item? Public Personnel Review, 19, 44-48.

Weiten, W. (1984). Violation of selected item construction principles in educational measurement. Journal of Experimental Education, 52, 174-178.

Wesman, A. G., & Bennett, G. K. (1946). The use of "none of these" as an option in test construction. Journal of Educational Psychology, 37, 541-549.

Wilber, P. H. (1966). Positional response set in the multiple choice examination. Dissertation Abstracts, 27, 2902-2903.

Williams, B. J., & Ebel, R. L. (1957). The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. The 14th Yearbook of the National Council on Measurement in Education, 63-65.

Williamson, [illegible] L., & Hopkins, K. D. (1967). The use of "none-of-these" versus homogeneous alternatives on multiple-choice tests: Experimental reliability and validity comparisons. Journal of Educational Measurement, 4, 53-58.

Zimmerman, W. S., & Humphreys, L. G. (1953). Item reliability as a function of the omission of misleads. American Psychologist, 8, 460-461.

APPENDIX A

REFERENCES--TEXTBOOKS' TREATMENT OF TEST ITEM WRITING

Adams, G. S. Measurement and evaluation in education, psychology, and guidance. New York: Holt, Rinehart, & Winston, 1964, pp. 339-342, 351, 368.

Ahmann, J. S. & Glock, M. D. Measuring and evaluating educational achievement (2nd ed.). Boston: Allyn and Bacon, 1975, pp. 65, 73-80, 109-111.

Blood, D. F. & Budd, W. Educational measurement and evaluation. New York: Harper and Row, 1973, pp. 81-94.

Brown, F. G. Measurement and evaluation. Peacock press, 1971, pp. 56-58.

Brown, F. G. Principles of Educational and Psychological Testing. Hinsdale, Illinois: Dryden Press, 1970.

Chase, C. I. Measurement for educational evaluation. Reading, MA: Addison Wesley, 1974, pp. 110-116.

Davis, F. B. Educational measurements and their interpretation. Belmont, CA.: Wadsworth, 1964, pp 267-284.

Denova, C. C. Test construction for training evaluation. New York: Van Nostrand Reinhold, 1979, pp. 51-65.

Downie, N. M. Fundamentals of measurement (2nd ed.). New York: Oxford Press, 1967, 149-167.

Durost, W. N. & Prescott, G. A. Essentials of measurement for teachers. New York: Harcourt, Brace & World, 1962,

Ebel, R. L. Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice Hall, Inc., 1972, pp. 135-162.

Furst, E. J. Constructing evaluation instruments. New York: Longmans, Green and Company, 1958.

Gerberich, J. R. Specimen objective test items: A guide to achievement test construction. New York: Longmans Green, 1956.

Green, J. A. Teacher-made tests. New York: Harper & Row, 1963, 32-38.

Green, J. A. Introduction to measurement and evaluation. New York, Dodd, Mead & Company, 1970, 203-206.

Gronlund, N. F. Measurement and evaluation in teaching (2nd ed.). New York: Macmillan, 1976, 132, 173-193.

Hawkes, H. E., Lindquist, E. F., & Mann, C. R. The construction and use of achievement examinations. Boston, Mass.: Houghton Mifflin, 1936.

Karmel, L. O. & Karmel, M. O. Measurement and evaluation in the schools (2nd ed.). New York: Macmillan, 1978, pp. 402-408.

Lindeman, R. H. Educational measurement. Glenview, Ill.: Scott Foresman, 1967, 78-83.

Lindvall, C. M. Testing and evaluation: An introduction. New York: Harcourt, Brace, and World, 1961, 69-73.

Lien, A. J., Measurement and Evaluation of Learning (4th ed.). Dubuque, IA: Brown, 1980, pp. 111-114.

Marshall, J. C. & Hales, L. W. Essentials of testing. Reading, MA: Addison-Wesley, 1972, 45-67.

Martuza, V. R. Applying norm-referenced and criterion-referenced measurement in education. Boston: Allyn and Bacon, 1977, 212-240.

Mehrens, W. A. & Lehmann, I. J. Measurement and evaluation in education and psychology (3rd ed.). New York: Holt, Rinehart, & Winston, 1984, 151-165.

Multiple-choice questions: A close look. Princeton, N. J.: Educational Testing Services, 1963.

Noll, V. H. Introduction to educational measurement (2nd ed.). Boston, MA.: Houghton Mifflin, 1965.

Nunnally, J. C. Educational measurement and evaluation (2nd ed.). New York: McGraw-Hill, 1972, 169-181.

Payne, D. A. The specification and measurement of learning outcomes. Waltham, MA: Blaisdell, 1968, 63-72

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978, 54-62.

Popham, W. J. Modern educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1981, 236, 251-262.

Roid, G. H. & Haladyna, T. M. A technology for test item writing. N. Y.: Academic Press, 1982.

Sax, G. Principles of educational and psychological measurement and evaluation (2nd ed.). Belmont, Ca.: Wadsworth, 1980, 101-113.

Stanley, J. C. & Hopkins, K. O. Educational and psychological measurement and evaluation (5th ed.). Englewood Cliffs, N. J.: Prentice-Hall, 1972, 232-255.

Tenbrink, T. D., Evaluation: A practical guide for teachers. New York: McGraw-Hill, 1974, 369-379.

Thorndike, R. L., & Hagen, E. Measurement and evaluation in psychology and education (3rd ed). New York: Wiley, 1969, 93-118.

Travers, R. M. How to make achievement tests. New York, N. Y.: Odyssey Press, 1950.

Wesman, A. G. Writing the test item. In R. L. Thorndike (Ed.) Educational Measurement. Washington, D. C.: American Council on Education, 1971, 99-116.

Wood, D. A. Test construction: Development and interpretation of achievement tests. Columbus, Ohio: C. E. Merrill, 1980.